1. **Provide your team background and organization description (if applicable).**

I am a Principal Data Scientist from a Canada AI startup company. This is my linkedin page: https://www.linkedin.com/in/ning-jia-04b55b241/

2. **Explain why you participated in the AllergenChip challenge.**

The challenge looks interesting. That's my first time tackling a multi label classification problem. Also, the dataset is tabular and relatively small. I like to deal with such dataset, because normally they won't require too many computing resources and training time.

3. **Describe how you built your winning model and elaborate on the technical and modeling choices you made.**

This is a journey of trails and experiments. After exploration, I found there are associations between the labels by performing association rules mining. So I intended to build one model that can capture all the labels.

Deep learning models in theory can be a good choice for this multi label classification task, but for this dataset they didn't perform well. For all single targets, tree-based algorithms outperform deep learning models. We can also train a MultiOutputClassifier model to predict all the labels. But the idea behind it is just fitting one model per target.

So I decided to train multiple binary classifiers for each target. We have 27 targets, so it's almost impossible to fine tune every model. I chose LightGBM and catboost, which typically perform well for small tabular dataset with categorical features. I use AUC as the metric for training because some targets are balanced, some are im-balanced, AUC generally can handle them well. I built a pipeline to automatically select features for each target and train multiple models with cross-validation and different seeds to reduce the randomness.

For the final submission, I selected the best threshold based on quantile and multiplier to get the best F1 score for the out-of-fold prediction. There are detailed descriptions in the submission notebook. Then the predictions for each target are ensembled results from weighted average predictions from LightGBM and catboost. Finally, the associations of the targets will be adjusted by the rules learned from targets association mining.

4. **Were the CPU/RAM resources provided in the challenge notebook sufficient from your point of view?**

The resources provided are sufficient for this competition.